Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Customers' abandonment strategy in an $M/G/1$ queue

Eliran Sherzer
Department of Industrial Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel, scherzter@gmail.com
Yoav Kerner
Department of Industrial Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel, kerneryo@bgu.ac.il

We consider an $M/G/1$ queue in which the customers, while waiting in line, may renege from it. We study the Nash equilibrium profile among customers, and show that it is defined by two sequences of thresholds. For each customer, the decision is based on the observed past (which determines from what sequence the threshold is taken), and the observed queue length (which determines the appropriate element in the chosen sequence). We construct the set of equations that has the Nash equilibrium as its solution, and discuss the relationships between the properties of the service time distribution and the properties of the Nash equilibrium, such as uniqueness and finiteness.

Key words: Stochastic Games/Group decisions, Markov processes, Probability

1. Introduction

Understanding customers' abandonments from a queue is of interest for both service providers and customers. There are many applications concerning customers' abandonments since real-world customers are unwilling to wait for excessive lengths of time. These applications were presented more broadly by Mandelbaum and Shimkin Mandelbaum and Shimkin (2000). Traditional queueing theory has dealt with the analysis of queues under the assumption of a given patience distribution, and many studies have addressed models which include abandonments in the literature, starting with Barrer Barrer (1957), who studied the queue length distribution in the $M/M/s+D^w$ case (where

D indicates deterministic patience). Sufficient conditions for the existence of the steady-state virtual waiting time distribution in the $G/G/1+G^w$ were later obtained by Baccelli Baccelli (1984, 1981). Boxma et al. Boxma et al. (2010) showed how to determine the busy period distribution for various choices of the patience time distribution, where the busy period is the waiting time in the queue. Brandt and Brandt Brant and Brant (2013) also studied the distribution of the busy period. In particular, they gave an explicit representation of the Laplace–Stieltjes transform of the workload and the busy period, in the case of phase-type distributed impatience. Yet, these studies assumed that the patience distribution was given. Here, we investigate the basis of these distributions, which are affected by factors such as individual costs and preferences. Essentially, the patience of each customer is based on an individual optimization, that is, the perceived balance between the costs of waiting and the benefits of service. Since the behaviour of others has an influence on the individual (abandonments of others shorten the individual's required waiting time), the mutual interactions lead us to look at the system in the standard form of a game, and to seek the Nash equilibrium. Therefore, the patience distribution is no longer given, but instead it results from a cost/reward model, and from the strategic behaviour implied by it.

The seminal study that viewed queues as economic systems, and studied the strategic behaviours within them, came from Naor Naor (1969). Naor presented the following model: an observable $M/M/1$ queue, in which there is a constant reward R from service, and a linear waiting cost C. He showed that under self-optimization, customers will join the queue if the number of customers present upon arrival is less than a threshold, n. This threshold is defined by $n=\lfloor\frac{R\mu}{C}\rfloor$. Moreover, once a customer joins the queue, he stays forever. Hassin and Haviv Hassin and Haviv (1995) also studied an $M/M/1$ queue, but in their case the customers have no information about their current position in the queue (i.e., unobservable queue). The waiting cost is the same as in Naor's study. However, the reward is $R>0$ if service is completed in less than T time units, otherwise it equals 0. They showed that the pure equilibrium is in the form of (p,τ). That is, customers join the queue with probability p and otherwise balk, and if they join then they wait τ time units.

Hassin and Haviv showed specifically that $\tau = T$. The reason for this result is that the virtual waiting time follows an increasing failure rate (IFR) pattern. That is, the remaining waiting time stochastically decreases along with the time passed. This is true for every $M/M/1$ queue with impatient customers, as Baccelli and Hebuterne previously showed in Baccelli (1981). Mandelbaum and Shimkin Mandelbaum and Shimkin (2000) retained the assumptions of unobservability, linear waiting costs and constant service reward, but instead considered an $M/M/m$ queue and heterogeneous customers whose waiting costs and rewards may vary between the different customer types. Their results indicate that, depending on the reward to cost ratio, a customer's best response was to either abandon the queue upon arrival unless of course one of the servers was available, in which case he enters service immediately or, given the IFR property, to never abandon and wait until receiving service. Of course, within a customer type, all customers follow the same strategy (either balk or stay forever), and therefore the Nash equilibrium is pure. However, a case in which the solution is richer is when customers are discharged without knowing it, (see, e.g. Palm (1953)). Mandelbaum and Shimkin proposed such a model. In their proposed model, denoted by $M/M/m(q)$, each customer will never be served with probability $1-q$. In this model, the longer a customer has already waited, the higher the posterior probability that he has been discharged. It turns out that the $M/M/m(q)$ system has an eventually decreasing (and in fact unimodal) hazard rate function, which makes finite abandonments rational. Thus, Mandelbaum and Shimkin showed that the best response is to either abandon at arrival, unless, of course, one of the servers is available or to abandon after a finite time T, which is determined by the ratio of the waiting cost and the reward. A follow up study by Mandelbaum and Shimkin Mandelbaum and Shimkin (2004) considered a nonlinear waiting cost in an unobservable $M/M/m$ queue. They provided conditions for the existence and uniqueness of the equilibrium, and suggested procedures for its computation. Moreover, they suggested a notion of equilibrium based on suboptimal decisions, the myopic equilibrium. Another important study, by Haviv and Ritov Haviv and Ritov (2001), explored an $M/M/1$ queue where the waiting cost is nonlinear and the reward from service is no

longer constant. They showed that the equilibrium strategy is to abandon after waiting T time units in the queue ($T = 0$ or $T = \infty$ are possible). Also, they showed that having a mixed Nash equilibrium may occur when the ratio of the cost and reward functions satisfies certain conditions. In this model, the waiting time has an IFR, and thus the remaining waiting time reduces with the elapsed waiting time. However, the waiting cost is convex, and therefore waiting becomes more expensive. The latter balances the improvement in the waiting time and the remaining waiting costs. So far, we have introduced a variety of models dealing with customers' abandonments from a queue. They vary in many ways, such as their observability level, their cost and reward functions, and number of servers. However, all of these studies had something in common, they all assumed an exponential service distribution. In this study, we relax this assumption. In particular, we consider an observable $M/G/1$ queue with a First-Come-First-Served (FCFS) discipline. With the term observable we mean that everyone can see their own position in the queue at any time, and they are able to observe service completions and abandonments made by others. Unlike the case where the service distribution is exponential, the age in case of generally distributed service time is meaningful. The age determines our anticipation of the remaining time for the current service, and consequently the waiting time. Hence, we allow customers to keep track of time. When considering utility, we specify that customers have a linear waiting cost and a constant reward from service. Customers don't have a waiting cost while they are being served. We assume that all customers are rational, in the sense that every customer wishes to maximize his own utility, and so they will stay as long as their utility is positive. Furthermore, we limit ourselves to service distributions with decreasing failure rate (DFR), that is, the remaining waiting time stochastically increases with time . In fact, in our model, we can use a weaker property. Specifically, we require only that the service time mean residual life (MRL) is increasing with its age. Consequently, customers' expected queueing time increases with the service age. If the service time MRL isn't bounded then rational customers will surely abandon at some point, since their utility will eventually be negative. An example of this kind of distribution is the Pareto distribution. However, if the service time MRL

is bounded, it is possible that customers will be willing to wait forever, given a big enough service reward. An example of a distribution of this kind is the hyperexponential. The rest of the paper is arranged as follows. In Section 2 we present the model, and customers' utility and strategy profile. Next, in Section 3 we give the theoretical framework with instructions on how to obtain the Nash equilibrium. This is followed by a numerical example. Finally, in Section 4 we discuss and summarize our primary results.

2. Model formulation

2.1. Model description

We consider a single server queueing model in which the arrivals are according to a Poisson process with rate λ and service times are iid generally distributed. We denote the service time by X, and $\mathbb{E}[X]$ by $\bar{x}$. We also use the notation $f(\cdot)$ for the pdf and $F(\cdot)$ for the cdf, with $\bar{F}(\cdot)=1-F(\cdot)$, and the hazard function is denoted by $h(\cdot)$. The service discipline is FCFS. Each customer can see his position in the queue at any moment, and he is able to observe service completions and abandonments made by others. However, customers are not aware of events that occurred prior to their arrival. Of course, they do not anticipate future service times. Customers can keep track of time, and are allowed to abandon at any moment. All customers are homonomous in their reward from service, which is denoted by V, and a linear waiting cost, which is denoted by C. In order to complete the model description we first present the following definition. For a non-negative random variable X with cdf $F(\cdot)$, the MRL function is defined as follows:

$$m_X(x)=\mathbb{E}[X-x|X>x]=\int_0^\infty \frac{\overline{F}(x+t)}{\overline{F}(x)}dt, \quad x\geq 0$$

Also, let $m_X(\infty)=\lim_{x\to\infty} m_X(x)$. We distinguish between service distributions in which the MRL is bounded, and those in which it is unbounded. We assume that the service time has an increasing mean residual life (IMRL) **?**.

2.2. Utility function

For each individual the utility function balances the reward from service with the expected waiting cost. Customers always take into account their future costs, while the time they already waited is considered a sunk cost. Each customer wishes to maximize his own utility; as a result, his best response is to stay as long as the utility is positive. To simplify, we first consider a case where abandonments are not allowed. Let $G_n(t)$ be the expected utility function value from staying until being served, for an individual that has n customers in front of him in the system when the current service age is t. Clearly,

$$G_n(t) = \begin{cases} V & n = 0 \\ V - C(\mathbb{E}[X - t|X > t] + (n-1)\bar{x}) & n \geq 1 \end{cases}$$

We next present a sequence of differential equations of the series $G_n(t)$ for $n \geq 1$.

Proposition 1 *The series $G_n(t)$ solves the sequence of differential equations $G_n'(t) = C - h(t)(G_{n-1}(0) - G_n(t))$ for $n \geq 1$.*

Proof First, one can see that $G_{n-1}(0) - G_n(t) = C\mathbb{E}[X - t|X > t]$ and hence the RHS of Proposition 1 equals $C - Ch(t)\mathbb{E}[X - t|X > t]$. We also note that $\mathbb{E}[X - t|X > t] = \frac{\int_{s=t}^{\infty} \bar{F}(s)ds}{\bar{F}(t)}$. Hence, the derivative of $G_n(t)$ is

$$\frac{dG_n(t)}{dt} = -C\frac{-\overline{F}(t)\overline{F}(t) + f(t)\int_{s=t}^{\infty} \bar{F}(s)ds}{\overline{F}(t)^2} = C - Ch(t)\mathbb{E}[X - t|X > t]$$

which matched the derivative presented in the proposition, and hence we complete the proof.

However, in our model, deriving the utility is not straightforward. Customers have to take into account the fact that abandonments may occur, that is, one needs to consider the possibility that he abandons a few moments later. Nonetheless, we next show that our utility function is similar to $G_n(t)$. This means that the possibility of abandoning later isn't reflected in the utility function. We distinguish between two types of customers. Type **i** customers observed service completion and type **ii** customers didn't. Due to the fact that keeping track of time is possible, type **i** customers

know the exact service age at any moment. However, since the current service age is unknown upon arrival, type **ii** customers can only estimate it. Let $U_n(t)$ be the expected utility of a type **i** customer taking the optimal action in the next moment, given that there are n others in front of him in the system and t time units elapsed since the last service completion.

Proposition 2

$$U_n(t) = (G_n(t))^+, \quad n \geq 0$$

Proof n=0 implies that an individual is already in service, and hence has no cost. Clearly, the optimal action in the next moment is to stay. For $n \geq 1$, we first prove the proposition for $n = 1$ and then prove for $n \geq 2$ by induction. For $n = 1$, we have

$$U_1(t) = (h(t)dtU_0(0) + (1 - h(t)dt)U_1(t + dt) - Cdt + o(dt))^+, \quad t \geq 0$$

If the RHS is negative, then the best response is to abandon, and hence the utility is zero. If the RHS is positive, then we have the following differential equation:

$$U_1'(t) = C - h(t)(U_0(0) - U_1(t))$$

where clearly, $U_0(0) = V$. This differential equation coincides with Proposition 1, and hence $U_1(t) = (V - C(\mathbb{E}[X - t|X > t] + \bar{x}))^+$. We first assume that $U_{n-1}(t) = (G_{n-1}(t))^+$. Based on our assumption, clearly, $U_{n-1}(0) = (V - C(n-1)\bar{x})^+$. Combining with

$$U_n(t) = h(t)dtU_{n-1}(0) + (1 - h(t)dt)U_n(t + dt) - Cdt + o(dt)$$

therefore

$$U_n'(t) = C - h(t)(U_{n-1}(0) - U_n(t))$$

and solving the differential equation complete the proof.

Let $\hat{U}_n(t)$ be the utility of a type **ii** customer taking the optimal action in the next moment, given that n other in front of him in the system and t time units elapsed since his arrival. Formally,

let N be the number of customers in the system and $A(t)$ be the service age, both upon arrival. Finally, let $X_n(t)$ follow the distribution of the service residual, given that upon the arrival of a tagged customer there were n customers in the system and t time units elapsed since then. That is, $X_n(t) \stackrel{d}{=} \{X - A(t) | N = n\}$.

Proposition 3

$$\hat{U}_n(t) = \begin{cases} V & n = 0 \\ (V - C(\mathbb{E}[X_n(t) - t | X_n(t) > t] + (n-1)\bar{x}))^+ & n \geq 1 \end{cases}$$

Proof For $n = 0$ there is no cost, similar to $U_n(t)$. For $n \geq 1$, we have

$$\hat{U}_n(t) = U_{n-1}(0)h_n(t)dt + (1 - h_n(t)dt)\hat{U}_n(t + dt) - Cdt + o(dt)$$

where $h_n(t)$ is the corresponding hazard function of $X_n(t)$. The differential equation is

$$\hat{U}_n'(t) = C - h_n(t)(\hat{U}_{n-1}(0) - \hat{U}_n(t))$$

For both $U_n(t)$ and $\hat{U}_n(t)$ there are differential equations with the same structure. Hence, their solutions also have the same structure, where in this case, X is replaced by $X_n(t)$.

REMARK 1. The fact that service time has an IMRL implies that both $U_n(t)$ and $\hat{U}_n(t)$ are decreasing with t.

2.3. Strategy profile

As mentioned, customers' best response differs depending on the number of customers in front of them and which of the two different customer types they belong to. Therefore, each pair of queue length and customer type needs to be considered separately. Also, customers may balk, and clearly when the queue is long enough customers will not join. This happens when the utility is negative from the moment one arrives. Hence, under the assumption of rationality of customers, there is a maximum number of customers in the system, and it is denoted by $n_{\max}$. We show how to obtain it

in section 3.2. The customers' strategy profile is as follows. All customers join if the system length is less than $n_{\max}$. Of course, joining customers will adapt their strategy according to the utility function, which is defined by the customer's type and the number of customers in front of them in the queue. Both $U_n(t)$ and $\hat{U}_n(t)$ are monotonically decreasing with t, and therefore the best response is unique and hence pure. Let T_n be the time a type **i** customer who has n customers in front of him in the system is willing to wait from the moment of service completion until the next service completion occurs, or abandonment is made by the customer in front of him. Let A_n be the time that a type **ii** customer who has n customers in front of him in the system is willing to wait from his arrival point until a service completion occurs, or abandonment is made by the customer in front of him. For any type of customer if, while waiting, service completion occurs, he updates his utility function and consequently his best response. However if, while waiting, the customer in front of him abandons, his best response is to abandon as well. This is because the customer in front, who abandoned first, gathered more information and his abandonment puts the customer behind in the exact same situation, to which the best response was to abandon. In conclusion, we have two sequences that define customers' strategies. The first one is $\{T_1, T_2, ..., T_{n_{\max}-2.}\}$ and the second is $\{A_1, A_2, ..., A_{n_{\max}-1}\}$. The largest index value of the second sequence is $A_{n_{\max}-1}$, because it is the largest value observed upon arrival for which one would be willing to join. That is, if a customer joined and observed $n_{\max}$ customers in the system, he would no longer join. In the first sequence, the largest index is obtained when a customer in the $n_{\max}$th position in the queue observes service completion. In this scenario, he has $n_{\max} - 2$ others in front of him after the departure of the customer who just completed his service. We next show how to obtain the values of T_n and A_n.

3. Results

3.1. Customer's best response

Using Propositions 2 and 3 we show how to obtain customers' best responses, for both type **i** and type **ii** customers, given $N = n$. Specifically, we show how to obtain the sequences $\{T_1, T_2, ..., T_{n_{\max}-2.}\}$ and $\{A_1, A_2, ..., A_{n_{\max}-1}\}$. We begin with type **i** customers.

Lemma 1 *If* $\lim_{t\to\infty} m_X(t) < \frac{V}{C} - (n-1)\bar{x}$*, and based on the utility function from Proposition 2,* T_n *is the value of* t *that solves*

$$V - C(\mathbb{E}[X-t|X>t] + (n-1)\bar{x}) = 0 \quad 1 \le n \le n_{\max}-2 \tag{1}$$

Otherwise, $T_n = \infty$*.*

Proof Since $n < n_{\max}-2$, the utility for $t=0$ is positive and of course, customers will be willing to wait as long their utility is positive. Recall that the service time follows IMRL. Also, the condition $\lim_{t\to\infty} m_X(t) < \frac{V}{C} - (n-1)\bar{x}$ means that for some value of t the utility will be negative. Therefore, Equation (1) has a unique and finite solution. Otherwise, the utility will be positive for every $t>0$ and hence the best response is to stay forever.

Of course, the solution of equation (1) is straightforward, and hence T_n for any possible n can be easily obtained.

Lemma 2 *If* $\lim_{t\to\infty} m_{X_n(t)}(t) < \frac{V}{C} - (n-1)\bar{x}$*, and based on the utility function from Proposition 3,* A_n *is the value of* t *that solves*

$$V - C(\mathbb{E}[X_n(t)-t|X_n(t)>t] + (n-1)\bar{x}) = 0, \quad 1 \le n \le n_{\max}-1 \tag{2}$$

Otherwise, $A_n = \infty$*.*

Proof We follow the same line of argument as Lemma 1.

Solving (2) is not straightforward, mainly because obtaining the distribution of $X_n(t)$ is challenging. As mentioned, $X_n(t) \stackrel{d}{=} \{X - A(t)|N{=}n\}$. That is, in order to obtain the distribution of $X_n(t)$, one must first obtain the distribution of $A(t)|N{=}n$. Let $R(t,a)$ follow the distribution of the residual service time, given that the service age upon arrival is a and t time units have elapsed since the customer's arrival. That is, $R(t,a) \stackrel{d}{=} X-(t+a)|X>(a+t)$. Therefore, by using the law of total probability, the following equation is equivalent to (2):

$$V - C\int_a (\mathbb{E}[R(t,a)|N=n] + (n-1)\bar{x}) f_{A(t)|N=n}(a)da = 0, \quad n \le n_{\max}-1 \tag{3}$$

Yet $f_{A(t)|N=n}(a)$ is unknown, and will be derived in the following sections.

3.2. Obtaining the maximum length of the queue

Proposition 4

$$n_{\max} = \sup\left\{ n \in \mathbb{N} : n \leq \frac{\frac{V}{C} - \int_a \mathbb{E}[R(0,a)|N=n] f_{A(0)|N=n}(a)da + \bar{x}}{\bar{x}} \right\} + 1 \tag{4}$$

Proof We seek the largest integer value of n that obeys $\hat{U}_n(0) > 0$. Thus, after extracting n from equation (3) and applying the supremum we get the expression in (4).

We observe that once $f_{A(0)|N=n}(a)$ is derived, $n_{\max}$ can be computed.

3.3. Markov chain underlying the process

Our motivation for using a Markov chain is mainly to obtain the pdf of the service age for a tagged customer who observes n others in the system upon arrival, with t time units having elapsed since then. This will eventually allow us to find a strategy that holds for the Nash equilibrium.

3.3.1. Markov chain state space First, we give some general notation for the steady states of the Markov chain:

$$S = \{(k, a, w_{k+1}, w_{k+2}, ..., w_{n-1})\}$$

where

- k is the number of waiting customers that observed service completion;
- a is the age of the current service;
- w_i is the waiting time of the i^{th} customer in the system; and
- n is the number of customers in the system.

A general steady-state density is denoted by $p(k, a, w_{k+1}, w_{k+2}, ..., w_{n-1})$, and the probability density of having n customers in the system and a service age of a is denoted by $\pi(n, a)$. It can be derived from the steady states of the Markov chain that

$$\pi(n,a) = \sum_{k=0}^{n-1} \int_{\underline{w}} p(k, a, w_{k+1}, ..., w_{n-1}) d\underline{w}, \quad n \geq 1,\ a \in \mathbb{R}^+,$$

We denote the marginal probability of having n customers in the system by π_n, which is derived as $\pi_n = \int_a \pi(n,a)da$.

We first indicate some general and rather trivial relationships. From the arrival order, we get $w_{n-1} < w_{n-2} .... < w_{k+1} < a$. Due to the abandonment strategies, we get $a < T_k$ for $1 \leq k \leq n_{\max}-2$. This is because, if $a > T_k$, then the k^{th} customer will have already abandoned by now. There are no constraints on a for $k=0$. That is, if no-one observed service completion, the first in the queue could arrive at any value of a. By the same argument,

$$w_i < A_i, \quad \text{for } k+1 \leq i \leq n-1.$$

For further analysis we present the following definition: let the *state structure* be the combination set of (n,k), where n is the number of customers in the system and k is the number of waiting customers that observed service completion. Since the queue length is limited, the Markov process has a limited number of different state structures.

Proposition 5 *The total amount of Markov chain state structures is*

$$|S_{(k,n)}| = \frac{n_{\max}(n_{\max}+1)}{2}$$

Proof We first claim that if there are n customers in system, then there are n different state structures for $1 \leq n \leq n_{\max}-1$. What determines the number of state structures for a given n, is the number of different possible values of k, where $0 \leq k \leq n-1$. That is, the values of k can be from zero to $n-1$, because even if everyone observed service completion there would still be one in service. However, for $n = n_{\max}$ there are $n_{\max}-1$ different state structures. In this case, $0 \leq k \leq n_{\max}-2$, while the state $\{n_{\max}-1, a\}$ is not possible. Finally, $n=0$ means an empty system, with just one state structure. The total number can be computed as an arithmetic progression. It is equivalent to the sum $\sum_{i=1}^{n_{\max}} i$, where each value of i represents the amount of state structures for a given n, except $i = n_{\max}$, which in this case includes cases for both n=0 and n=$n_{\max}$. From here the result is straightforward.

3.3.2. Finding the steady-state densities Due to the complexity of the process we begin with a simple example. Let us consider state $(0,a)$, which refers to an active server with a current service age a and an empty queue. For $a<A_1$,

$$p(0,a)=p(0,0)e^{-\lambda a}\bar{F}(a) \tag{5}$$

Equation (5) justifies the following. State $(0,a)$ will always follow state $(0,0)$. This means that state $(0,0)$ occurred, and during the intervening a time units there were no service completions and no arrivals. Thus, the probability density of $p(0,a)$ is as for $p(0,0)$, times the probability that there were no service completions nor arrivals. However, for $a>A_1$, we allow arrivals to occur from the beginning of service, assuming that they will have abandoned by the time the service reaches age a. For $mA_1\leq a<(m+1)A_1$, where $m\in\{0,1,2,3...\}$, it is possible that a customer will arrive and abandon after A_1 time units. If, during the stay of the new arrival, more customers arrive, then they will not be in the queue once he abandons. This is because, if they stayed until then, they would abandon as well, since the customer ahead of them abandoned. This process can happen no more than m times, for $a<(m+1)A_1$. For example, if $m=1$, which means $A_1<a<2A_1$, then two cases are possible: no arrivals at all and no service completion, or one arrival who abandoned before the state reaches $(0,a)$ and no service completion. From basic probability we obtain

$$p(0,a)=p(0,0)(e^{-\lambda a}+\lambda e^{-\lambda a}(a-A_1))\bar{F}(a)$$

Thus, $p(0,a)$ is equal to $p(0,0)$ times the probability that no service completion occurs and there were from 0 to m arrival events followed by abandonments. We next give a general expression. Let $g(k,n,m,a)$ be

$$g(k,n,m,a)=\begin{cases}\left(\sum_{j=0}^{m}\frac{\lambda^j e^{-\lambda(a-jA_n)}(a-jA_n)^j}{j!}\right)\overline{F}(a) & 1\leq k+1=n\leq n_{\max}-2\\ \left(\sum_{j=0}^{m}\frac{\lambda^j e^{-\lambda(w_{n-1}-jA_n)}(w_{n-1}-jA_n)^j}{j!}\right)\frac{\overline{F}(a)}{\overline{F}(a-w_{n-1})} & 1\leq k+1<n\leq n_{\max}-1\\ \frac{\overline{F}(a)}{\overline{F}(a-w_{n-1})} & n=n_{\max}\end{cases}$$

and

$$m \in \begin{cases} \mathbb{N}_0 & k=0 \\ \{0,1,2,...\left\lfloor \frac{T_k - A_n}{A_n} \right\rfloor\} & 1 \le k \le n_{\max}-2 \end{cases}$$

where a is the current service age, w_{n-1} represents the waiting time of the last joining type **ii** customer who observed $n-1$ customers in the system upon arrival. n is the number of customers in the system, and m relates to the possible values of a: specifically, $mA_n < a < (m+1)A_n$. Lastly, k is the number of customers who observed service completion.

Lemma 3 *The function $g(k,n,m,a)$ is represented differently in three cases.*

Case 1, with $1 \le k+1=n \le n_{\max}-2$, represents the probability that the current state is (k,a), given that a time units ago the state was $(k,0)$.

Case 2, with $1 \le k+1 < n \le n_{\max}-1$, represents the probability that the current state is now $(k,a,w_{k+1}...w_{n-2},w_{n-1})$, given that w_{n-1} time units ago the state was $(k,a-w_{n-1},w_{k+1}-w_{n-1},...,w_{n-2}-w_{n-1},0)$.

Case 3 represents the probability that the current Markov state is $(k,a,w_{k+1}...,w_{n_{\max}-2},w_{n_{\max}-1})$, given that $w_{n_{\max}-1}$ time units ago the Markov state was:

$(k,a-w_{n_{\max}-1},w_{k+1}-w_{n_{\max}-1},...,w_{n_{\max}-2}-w_{n_{\max}-1},0)$.

Proof Case 1: in order that the state $(k,0)$ will be replaced by the state (k,a) after a time units, we need to ensure that there will not be a service completion during those a time units. Also, we need to ensure that there will not be any new arrivals, or if there are, then they will have abandoned by the time the Markov chain state reaches (k,a). The probability of no service completion is $\bar{F}(a)$. The probability of not having new arrivals once the state reaches (k,a) is

$$\sum_{j=0}^{m} \frac{\lambda^j e^{-\lambda(a-jA_n)}(a-jA_n)^j}{j!}$$

Of course, scenarios which include more than one abandonment made by the $(n+1)^{\text{th}}$ customer in the system are under consideration, where j is the number of times it occurs. We also note that

$j \leq m$.

Case 2: in order that the state $(k, a - w_{n-1}, w_{k+1} - w_{n-1}, ..., w_{n-2} - w_{n-1}, 0)$ will be replaced by the state $(k, a, w_{k+1}...w_{n-2}, w_{n-1})$ after w_{n-1} time units, we need to ensure that there will not be service completion and no new arrivals which stayed during that time (similar to Case 1). The probability that there will not be service completion is $\mathbb{P}(X > a|X > a - w_{n-1})$, which is equivalent to $\frac{\bar{F}(a)}{\bar{F}(a-w_{n-1})}$. In Case 3, new arrivals are not a possibility anyway, and therefore, in order that the state will be transposed from $(k, a - w_{n_{\max}-1}, w_{k+1} - w_{n_{\max}-1}, ..., w_{n_{\max}-2} - w_{n_{\max}-1}, 0)$ to $(k, a, w_{k+1}..., w_{n_{\max}-2}, w_{n_{\max}-1})$, we only need to ensure that there will be no service completion, which is $\frac{\bar{F}(a)}{\bar{F}(a-w_n)}$ as in Case 2.

Hence, from Lemma 3 we have, for $mA_n \leq a \leq (m+1)A_n$,

$$p(k, a, w_{k+1}, ..., w_{n-1}) = p(k, a - w_{n-1}, ..., w_{n-2} - w_{n-1}, 0)g(k, n, m, a) \tag{6}$$

From (6) we see that it is possible to separate the expression of the steady-state densities of the Markov chain into two parts. The first one is also a steady-state density of the Markov chain, for which the last argument is set to be 0. The second is the function $g(k, n, m, a)$, which is computable given the model parameters. Therefore, in order to obtain the steady-state densities of the Markov process we need to find those in which the last argument is set to be 0. From the balance equations,

$$\lambda\pi_0 = \int_a p(0, a)h(a)da \tag{7}$$

$$\lambda p(k, a, w_{k+1}, ..., w_{n-1}) = p(k, a, w_{k+1}, ..., w_{n-1}, 0) \tag{8}$$

$$\int_a \sum_{i=0}^{n-1} p(i, a, w_{i+1}, ..., w_{n+1})h(a)da = p(n, 0) \tag{9}$$

Recall from Proposition 5 that the number of state structures is $\frac{n_{\max}(n_{\max}+1)}{2}$. Excluding the state 0 for each state structure, there is single state for which the argument is 0. Therefore, from equations (7) to (9) we have $\frac{n_{\max}(n_{\max}+1)}{2}$ equations. Where in fact from (7) consist one Equation, (8) consist $\frac{n_{\max}(n_{\max}-1)}{2} + 2$ and (9) consist $n_{\max} - 3$ equations. Combined with (6) and the fact that $\sum_{n=0}^{n_{\max}} \pi_n = 1$, all the steady states of the Markov process is derived.

REMARK 2. For numerical computations, we guess a value for π_0. Using equations (6) to (9), we compute $\sum_{n=0}^{n_{\max}} \pi_n$. If the total sum is smaller than 1, we guess a larger number for π_0 and vice versa.

We next give a special case where $n_{\max}=3$. There are 6 different state structures, and they are represented as $\{(0),(0,a),(0,a,w_1),(1,a),(1,a,w_2),(0,a,w_1,w_2)\}$. From balance equations we state that

$$\lambda\pi_0 = \int_{a=0}^{\infty} p(0,a)h(a)da \tag{10}$$

Also,

$$\lambda p(0,a) = p(0,a,0) \tag{11}$$

$$\lambda p(0,a,w_1) = p(0,a,w_1,0) \tag{12}$$

$$\int_{a=0}^{\infty} p(0,a,w_1)h(a)da = p(0,0) \tag{13}$$

$$\int_{a=0}^{\infty} p(0,a,w_1,w_2)h(a)da = p(1,0) \tag{14}$$

$$\lambda p(1,a) = p(1,a,0) \tag{15}$$

Using equations (6) and (10) to (15), and applying the numerical procedure from Remark 2, all steady states can be computed.

3.4. The age distribution given the queue length

We next show how to obtain $f_{A(t)|N=n}(a)$, while using the steady-state probability densities of the Markov chain. Let Y follow the distribution of the total amount of time a tagged customer waited from arrival until either he abandons or there was service completion, sampled by an outside inspector at an arbitrary moment. In fact, $\{A(y)|N=n\} \stackrel{d}{=} \{A|N=n,Y=y\}$. From Bayes' law,

$$f_{A|N=n,Y=y}(a) = \frac{f_{Y|N=n,A=a}(y)f_{A|N=n}(a)}{\int_a f_{Y|N=n,A=a}(y)f_{A|N=n}(a)da}. \tag{16}$$

Since the presentation in a general case is implicit, we demonstrate using a special case of $n_{\max}=3$. However, we can proceed similarly for any value of $n_{\max}$. We show separately how to obtain $f_{A|N=n}(a)$ and $f_{Y|N=n,A=a}(y)$ for both $N=1$ and $N=2$. $f_{A|N=n}(a)$ can be derived directly from the steady-state densities of the Markov process, specifically for $N=1$,

$$f_{A|N=1}(a)=\frac{p(0,a)}{\pi_1}$$

and for $N=2$,

$$f_{A|N=2}(a)=\frac{p(1,a)+\int_{w_1=0}^{A_1\wedge a}p(0,a,w_1)dw_1}{\pi_2}$$

We next derive $f_{Y|N=1,A=a}(y)$. Let Q be a random variable that represents the inter-arrival times. Of course, $Q\sim \text{Exp}(\lambda)$. Due to the PASTA property, an outside inspector sampling times is equivalent to customer arrival times. Therefore, $Q|Q\le R(0,a)\wedge A_1$ is equivalent to $Y|N=1,A=a$.

Lemma 4 *The conditional density of Y given $A=a$, $N=1$ is*

$$\frac{\lambda e^{-\lambda y}\frac{\bar{F}(a+y)}{\bar{F}(a)}}{\mathbb{P}(Q\le A_1\wedge\ R(0,a))}$$

and

$$\mathbb{P}(Q\le A_1\wedge\ R(0,a))=\int_{r=0}^{A_1}(1-e^{-\lambda r})dr+\int_{r=A_1}^{\infty}(1-e^{-\lambda A_1})dr$$

Proof

$$\mathbb{P}(Q\le y|Q\le R(0,a)\wedge A_1)=\frac{\mathbb{P}(Q\le y,Q\le R(0,a)\wedge A_1)}{\mathbb{P}(Q\le R(0,a)\wedge A_1)}$$

We give explicit expressions for both numerator and denominator.

The numerator is

$$\mathbb{P}(Q\le y,Q\le R(0,a)\wedge A_1)=\int_{r=0}^{y}\mathbb{P}(Q\le r)\frac{f(a+r)}{\bar{F}(a)}dr+\int_{r=y}^{\infty}\mathbb{P}(Q\le y)\frac{f(a+r)}{\bar{F}(a)}dr$$

and the denominator is

$$\mathbb{P}(Q\le R(0,a)\wedge A_1)=\int_{r=0}^{A_1}\mathbb{P}(Q\le r)\frac{f(a+r)}{\bar{F}(a)}dr+\int_{r=A_1}^{\infty}\mathbb{P}(Q\le A_1)\frac{f(a+r)}{\bar{F}(a)}dr$$

After taking the derivative, we get the pdf of $Q|Q\le R(0,a)\wedge A_1$, and hence of $Y|N=1,A=a$ as well.

We next derive $f_{Y|N=2,A=a}(y)$. Let W_1 be a random variable taking value of w_1 given that the state is $(0,a,w_1)$, sampled at an arbitrary moment. The pdf of W_1 is denoted by $f_{W_1}(w_1)$, and it is equivalent to $\frac{p(0,a,w_1)}{\int_{u=0}^{A_1 \wedge a} p(0,a,u)du}$. When a customer arrives to a system given that $N=2$, there are two possible state structures: **i**. $p(1,a)$ and **ii**. $p(0,a,w_1)$. Let I be an indicator that receives the value of 1 when the state structure is $p(1,a)$, and receives 0 when it is $p(0,a,w_1)$. From simple probability considerations, $\mathbb{P}(I=1) = \frac{p(1,a)}{p(1,a)+\int_{w_1=0}^{A_1 \wedge a} p(0,a,w_1)dw_1}$. We indicate that $Q|Q \leq R(0,a) \wedge A_2 \wedge (I(T_1 - a) + (1-I)(A_1 - W_1))$ is equivalent to $Y|N{=}2,A{=}a$.

Lemma 5 *The conditional density of Y given $A=a$, $N=2$ and $Y < (I(T_1-a)+(1-I)(A_1-W_1)) \wedge A_2 \wedge R(0,a)$ is*

$$\mathbb{P}(I=1)\frac{\lambda e^{-\lambda y}\frac{\bar{F}(a+y)}{\bar{F}(a)}}{\mathbb{P}(Q \leq A_2 \wedge R(0,a) \wedge (T_1-a))} + \mathbb{P}(I=0)\int_{w_1=0}^{A1}\frac{\lambda e^{-\lambda y}\frac{\bar{F}(a+y)}{\bar{F}(a)}}{\mathbb{P}(Q \leq A_2 \wedge R(0,a) \wedge (A_1-w_1))}p(0,a,w_1)dw_1$$

and

$$\mathbb{P}(Q \leq A_2 \wedge R(0,a) \wedge (T_1-a)) = \int_{r=0}^{(T_1-a)\wedge A_2}\mathbb{P}(Q \leq r)\frac{f(a+r)}{1-F(a)}dr + \int_{r=A_2\wedge(T_1-a)}^{\infty}\mathbb{P}(Q \leq (T_1-a)\wedge A_2)\frac{f(a+r)}{\bar{F}(a)}dr$$

and

$$\mathbb{P}(Q \leq A_2 \wedge R(0,a) \wedge (A_1-w_1)) = \int_{r=0}^{(A_1-w_1)\wedge A_2}\mathbb{P}(Q \leq r)\frac{f(a+r)}{\bar{F}(a)}dr + \int_{r=(A_1-w_1)\wedge A_2}^{\infty}\mathbb{P}(Q \leq (A_1-w_1)\wedge A_2)\frac{f(a+r)}{\bar{F}(a)}dr$$

The proof is given in Appendix A.

3.5. Relations between the thresholds

In this section we refer to some trivial and nontrivial results concerning the values of the sequences $\{A_1, A_2, ..., A_{n_{\max}-1}\}$ and $\{T_1, T_2, ..., T_{n_{\max}-2}\}$. Specifically, we describe the dependencies and boundaries between them. Intuitively, the more customers are in front of you, the less you would be willing to wait.

Lemma 6 *If* $T_i < \infty$, $T_i > T_{i+1}$ *for* $1 \leq i \leq n_{\max}-3$.

Proof Recall that t in equation (1) refers to the time one waited in the queue since service completion. Due to the fact that $\mathbb{E}[X-t|X>t]$ is increasing with t, it follows straightforwardly that the larger the value of n the lower the value of t that solves the equation, and hence the lower the value of T_n.

Lemma 7 *If* $A_i < \infty$, $A_i > A_{i+1}$ *for* $1 \leq i \leq n_{\max}-2$.

Proof Suppose an individual observed $n+1$ customers upon arrival. The last event prior to his arrival could be either an arrival, an abandonment or a service completion. If it were an arrival, then clearly the one who observed n customers was in a better situation when he arrived. If it was an abandonment, then this individual (if he knew) should not join at all, and hence he is in a worse situation than the one who observed n upon arrival. The last case is that where the arriving customer is the first to arrive during the current service period. In this case, the age of the service time is his inter-arrival time. Yet, the greater the queue length, the smaller the probability of such an event.

We also observe that all $\{T_1, T_2, ...T_{n_{\max}-2}\}$ are obtained independently of everything else. This is a direct result from equation (1), where the value of T_n is determined by the values of the model parameter and n. Moreover, we claim that A_i, for $\forall i \in \{1, 2, ...n_{\max}-2\}$, depends only on $\{A_1, A_2, ..., A_{i-1}\}$ and $\{T_1, T_2, ..., T_i\}$, and for $i = n_{\max}-1$, is dependent only on $\{A_1, ...A_{i-1}\}$ and $\{T_1, T_2, ..., T_{n_{\max}-2}\}$. This result is rather surprising, because intuitively the entire history of a busy period would effect the value of A_i. But the values of A_i are not affected by $\{A_{i+1}, ..., A_{n_{\max}-1}\}$ and $\{T_{i+1}, ...T_{n_{\max}-2}\}$. The intuition behind this is that the queue state upon arrival of one who observed i customers is not affected in any way by customers that arrived and observed $i+1$ or more in the current busy period. This is due to the fact that customers don't make their decisions with respect to what happens in the queue behind them. Customers are only interested in what happens in front of them. This is what allows us to compute the values of A_i for $\forall i \in \{1, 2, ..., n_{\max}-1\}$.

Theorem 1 *The Nash equilibrium profile is defined by two finite sequences of thresholds,* $\{T_1, T_2, ...T_{n_{\max}-2}\}$ *and* $\{A_1, A_2, ...A_{n_{\max}-1}\}$*, which are distinguished by the customer types. Within each sequence, from an individual point of view, each threshold is determined by the number of customers in front of him.*

Proof Customers' strategies are determined by their utility functions. Based on Propositions 2 and 3, we claim that there are two sequences because the different customer types have different utility functions, and therefore their strategies are different as well. The sequence lengths are a direct result of Proposition 4 and the definition of the steady states of the Markov chain. Finally, the fact that the Nash equilibrium profile within each sequence is defined by thresholds was already proved in Lemma 1 and Lemma 2.

3.6. Numerical result

We present an example where the service distribution is hyperexponential, and the model parameters are $V = 4.85$, $c = 1$, $\mu_1 = 1$, $\mu_2 = 0.2$, $p = 0.95$ and $\lambda = 3$. The (symmetric) Nash equilibrium is $n_{\max} = 3$, $T_1 = 7.73$, $A_1 = 7.202$ and $A_2 = 3.13$.

4. Conclusions

In this study we show how to obtain the Nash equilibrium in an observable $M/G/1$ queue with abandonments. We focus on service time distributions which have an IMRL. The Nash equilibrium is defined by two sequences of thresholds. The values of the sequence $\{T_1, T_2, ...T_{n_{\max}-2}\}$ are obtained by solving a linear equation. However, obtaining the values from the sequence $\{A_1, A_2, ...A_{n_{\max}-1}\}$ is much more difficult. They can be computed recursively based on the definition of the Markov process. The reason we are able to obtain them recursively is because customers' decisions are not effected by future arrivals, and therefore they are transparent to them. In other words, from the point of view of a customer who is in the n^{th} position in the queue, the maximum length of the queue is n. Finally, a numerical example is given in which both sequences are computed.

Acknowledgments

Helpful comments made by Moshe Haviv and Joseph Kreimer are highly appreciated. This research is supported by the Israel Science Foundation, grant No. 1319/11.

References

Baccelli F, Boyer F, Hebuterne G (1984) Single server queues with impatient customers. *Appl Probability*, 16:887–905.

Baccelli F, Hebuterne G (1981) On queues with impatient customers. *Performance*, 159–179.

Barlow R.E, Proschan F (1975) Statistical theory reliability and life testing: Probability models. *Hold, Rinehart and Winston*, New York.

Barrer D.Y (1957) Queuing with impatient customers and ordered service. *Oper. Res*, 5:650–656.

Boxma O, Perry D, Stadje W, Zacks S (2010) The busy period of an M/G/1 queue with customer impatience. *J. Appl. Probab*, 45(1):130–145.

Brandt A, Brandt M (2013) Workload and busy period for the M/GI/1 with a general impatience mechanism. *Queueing Systems*, 75:189–209.

Hassin R, Haviv M (1995) Equilibrium strategies for queues with impatient customers. *Oper. Res. Lett*, 17:41–45.

Haviv M, Ritov Y (2001) Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing Systems*, 38:495–508.

Mandelbaum A, Shimkin N, (2000) A model for rational abandonments from invisible queues. *Queueing Systems*, 36:141–173.

Mandelbaum A, Shimkin N (2004) Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems*, 47:117–146.

Naor P (1969). The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24.

Palm C (1953). Methods of judging the annoyance caused by congestion. *Tele*, 2:1–20.

Appendices

A. Appendix A

Proof

$$\begin{aligned}&\mathbb{P}(Q\le y|Q\le R(0,a)\wedge A_2\wedge(I(T_1-a)+(1-I)(A_1-W_1)))\\&=\frac{\mathbb{P}(Q\le y,Q\le R(0,a)\wedge A_2\wedge(I(T_1-a)+(1-I)(A_1-W_1)))}{\mathbb{P}(Q\le R(0,a)\wedge A_2\wedge(I(T_1-a)+(1-I)(A_1-W_1)))}\end{aligned}$$

We give explicit expressions for both the numerator and the denominator.

The numerator is

$$\begin{aligned}&\mathbb{P}(Q\le y,Q\le R(0,a)\wedge A_2\wedge(I(T_1-a)+(1-I)(A_1-W_1)))\\&=\int_{w_1=0}^{A_1\wedge a}\mathbb{P}(I=1)\mathbb{P}(Q\le y,Q\le R(0,a)\wedge A_2\wedge(T_1-a))\\&+\mathbb{P}(I=0)\int_{a=0}^{A_2\wedge(A_1-w_1)}\mathbb{P}(Q\le y,Q\le R(0,a)\wedge A_2\wedge(A_1-w_1))p(0,a,w_1)da\end{aligned}$$

and the denominator is

$$\begin{aligned}&\mathbb{P}(Q\le R(0,a)\wedge A_2\wedge(I(T_1-a)+(1-I)(A_1-W_1)))\\&=\int_{w_1=0}^{A_1\wedge a}[\mathbb{P}(I=1)P(Q\le R(0,a)\wedge A_2\wedge(T_1-a))\\&+\mathbb{P}(I=0)\mathbb{P}(Q\le R(0,a)\wedge A_2\wedge(A_1-w_1))]p(0,a,w_1)dw_1\end{aligned}$$

After taking the derivative, we obtain the pdf of $Q|Q\le R(0,a)\wedge A_2\wedge(I(T_1-a)+(1-I)(A_1-W_1))$, and hence of $Y|N=2,A=a$ as well.